DOCUMENT RESUME

ED 091 721 CS 201 326

AUTHOR	Smith, Vernon H.
TITLE	Composition Rating Scale.
PUB DATE	66
NOTE	19p.; Reprinted from "Research in the Teaching of English," Fall, 1969; See related documents CS 201 320-375
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Composition (Literary); *Educational Research; English Instruction; *Evaluation Methods; Higher Education; Language Arts; *Measurement Instruments; Research Tools; Resource Materials
IDENTIFIERS	Composition Rating Scale; *The Research Instruments Project; TRIP

ABSTRACT

Designed to assess consistency in teacher judgment of student essays and to assess conformity of teacher judgment with expert judgment, the Composition Rating Scale (CRS) requires the taker to rank-order five brief compositions. Requiring twenty minutes to complete, the scale can be used to evaluate the consistency of teacher judgments of compositions, to screen lay-composition readers, or to prepare student teachers. [This document is one of those reviewed in The Research Instruments Project (TRIP) monograph "Measures for Research and Evaluation in the English Language Arts" to be published by the Committee on Research of the National Council of Teachers of English in cooperation with the ERIC Clearinghouse on Reading and Communication Skills. A TRIP review which precedes the document lists its category (Teacher Competency), title, author, and date, and describes the instrument's purpose and physical characteristics.] (RB)

NCTE Committee on Research

The Research Instruments Project (TRIP)

ED 091721

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The attached document contains one of the measures reviewed in the TRIP committee monograph titled:

Measures for Research and Evaluation in the English Language Arts

TRIP is an acronym which signifies an effort to abstract and make readily available measures for research and evaluation in the English language arts. These measures relate to language development, listening, literature, reading, standard English as a second language or dialect, teacher competencies, or writing. In order to make these instruments more readily available, the ERIC Clearinghouse on Reading and Communication Skills has supported the TRIP committee sponsored by the Committee on Research of the National Council of Teachers of English and has processed the material into the ERIC system. The ERIC Clearinghouse accession numbers that encompass most of these documents are CS 201320 -CS 201375.

TRIP Committee:

W.T. Fagan, Chairman
University of Alberta, Edmonton

Charles R. Cooper
State University of New York
at Buffalo

Julie M. Jensen
The University of Texas at Austin

Bernard O'Donnell
Director, ERIC/RCS

Roy C. O'Donnell
The University of Georgia
Liaison to NCTE Committee
on Research

CS 201376

NATIONAL COUNCIL OF TEACHERS OF ENGLISH
1111 KENYON ROAD
URBANA, ILLINOIS 61801

Category: Teacher Competency

Title: Composition Rating Scale

Author: Vernon H. Smith

Description of the Instrument:

Purpose: To assess consistency in teacher judgment of essays and to assess conformity of teacher judgment with expert judgment.

Date of Construction: 1966

Physical Description: The CRS requires the taker to rank-order five brief compositions. A simple and efficient scoring scheme is based on deviation from experts' ranking of the same compositions. The test has two forms.

Requiring twenty minutes to complete, it could be used in studies where evaluating the consistency of teacher judgment of compositions is important. It could also be used to screen lay-composition-reader applicants. With the outside criterion (experts' rankings) and with the ease of comparing judgments within a teacher group, it could be useful for teacher training.

Validity, Reliability, and Normative Data:

The best evidence offered by the author for the validity of the test is the high degree of agreement among the experts who determined the final ranking (for scoring purposes) of the essays. Interrater reliabilities were .92 and .85 for two administrations of Form A and .88 and .84 for two administrations of Form B. The test-retest reliability of the experts on each form was 1.00.

The basic validity question, of course, is whether the teachers' judgment and ranking of the five test compositions is very similar to the judgments they make on actual compositions. No evidence is reported

on that. Since the test compositions are limited to only one kind of writing--a brief, personal letter in narrative form to a pen pal--the test does not assess teacher judgment of other kinds of writing.

The reliability coefficient from scores on both forms by teachers was .61. The test-retest reliability was .74 and .79 for Forms A and B respectively. When the two forms were considered together as a larger ten-item test, the test-retest reliability rose to .87. The author concludes that "the most reliable results will be obtained when the two forms of the test are given at the same time and the scores on each are combined to give a total score."

Ordering Information:

EDRS

Related documents:

More information (and Form A) of the test is available in Vernor H. Smith, "Measuring Teacher Judgment in the Evaluation of Written Composition," Research in the Teaching of English, 3 (Fall 1969), 181-195.

Thomas E. Whalen, "A Validation of the Smith Test for Measuring Teacher Judgment of Written Composition," Education, 93, No. 2, 172-175.

Vernon H. Smith
"Composition Rating Scale"
1966

RESEARCH IN THE
TEACHING OF ENGLISH

VOLUME 3, NUMBER 2, FALL 1969

Contents

ARTICLES

127 Drama in the secondary school:
A study of objectives
by James Hoetker and Richard Robb

160 Understanding allusions in literature
by Marjorie Roberts

166 The effects of prereading assistance
on the comprehension and attitudes
of good and poor readers
by Richard J. Smith and Karl D. Hesse

178 ITA and TO training in the development of children's creative writing
by Joanne A. Auguste and Fredric B. Nalven

181 Measuring teacher judgment in
the evaluation of written composition
by Vernon H. Smith

196 Teaching punctuation in the ninth grade
by means of intonation cues
by Jeanette R. Held

209 An experimental comparison of writing
achievement in English composition
and humanities classes
by William C. Budd

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
National Council of Teachers of English
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

Mr. Smith explains how he developed and validated two forms of a composition scale to test teacher evaluation of elementary writing. After trying out the test on various groups of prospective, beginning, and experienced teachers, he concludes that valid judgment of the quality of elementary writing is independent of experience, academic preparation, and professional training.

Measuring teacher judgment in the evaluation of written composition[1]

VERNON H. SMITH
Indiana University

We have known for a long time that raters would not always agree on the value of a particular composition.

> It is common knowledge to student and teacher alike that the grading of essay materials can be highly inconsistent. The grade given to an English theme may vary considerably among different raters and even with the same rater at different times.[2]

Perhaps the most dramatic evidence of this lack of agreement came from the study by Diederich, French, and Carlton.[3] In *Research in Written Composition* this study is summarized as follows:

---

[1] This article is based on a doctoral study completed in 1965 under the direction of Professor Harold M. Anderson at the University of Colorado and on subsequent research still in progress by the author. The dissertation, titled *An investigation of teacher judgment in the evaluation of written composition including the development of a test for the measurement thereof*, is available from University Microfilms, Ann Arbor, Michigan (Order No. 67-10, 011).

[2] J. C. Follman and J. A. Anderson, "An investigation of the reliabil-

> They (Diederich, French, and Carlton) analyzed the way ten English teachers rated 300 two-hour compositions by college freshmen in comparison to 43 other raters: social scientists, natural scientists, writers and editors, lawyers, and business executives. The raters were given no standards or criteria for judging the papers, merely asked to sort the themes into nine piles in order of general merit, with not less than 4 per cent of the papers in any pile. It was "disturbing to find that 94 per cent of the papers received either seven, eight, or nine of the nine possible grades, that no paper received less than five different grades, and that the median correlation between readers was .31. Readers in each field, however, agreed slightly better with the English teachers than with one another."[4]

While a number of studies have shown similar disagreement among theme raters, the Diederich-French-Carlton research is impressive because of the number of raters, the number of themes, and the magnitude of disagreement. The themes in this study were written by college freshmen. Other studies in this area have focussed on themes written by high school or college students. Studies below the high school level are rare.

Paul Diederich once summed up his investigations of the rating of compositions by English teachers as follows:

> The average commentary on teachers' comments need not be quite so brutal as this, for I have compressed into one paragraph a large number of flaws that I have found in many samples of papers marked by teachers that I have examined in research studies. I hate to say it, for I am kindly disposed toward all English teachers, but the dominant impression left by these studies is that the average English teacher, both in high school and in freshman composition courses, is barely literate, *capricious in judgment*, full of prejudices that have no basis in anyone's system of grammar, rhetoric, or style, hard to decipher, eager to misinterpret, and given to comments that have no connection with anything the student has written. . . .[5] (italics mine)

---

ity of five procedures for grading English themes," *Research in the Teaching of English*, 1967, *1*, 190.

[3] P. B. Diederich, J. W. French, and S. T. Carlton, *Factors in judgments of writing ability* (Research Bulletin RB61-15, Princeton, N.J.: ETS, 1961).

[4] R. Braddock, R. Lloyd-Jones, and L. Schoer, *Research in written composition* (Champaign, Ill.: NCTE, 1963), p. 41.

[5] P. B. Diederich, "The problem of grading essays" (Princeton, N.J.: ETS, 1957, pp. 7-8. Mimeographed.).

Capricious Rating

When the Diederich-French-Carlton report was published in 1961, I was the K-12 supervisor of English in a large suburban school district. I wondered whether the same "capricious judgment" found among high school and college English teachers could exist among teachers at lower grade levels. It might be very difficult for a third grader, or a fifth grader, or an eight grader to develop his composition skills if his teacher rated his themes high one year and if another teacher rated them low the next year. Is it possible for a student to get a teacher whose judgment is contrary to that of most other teachers?

Armed with curiosity and some themes I had borrowed from a fifth grade teacher, I attended a meeting of teachers from grades one through nine. Since the purpose of the meeting was to discuss the teaching of composition, the teachers were willing to participate in the little exercise I had prepared. Each teacher was given copies of the same seven short themes and asked to pick the two best and two worst. The fifth grade teacher and I had already picked a couple we thought were good, a couple we thought were poor and three that were in between. Although the majority of the teachers at the meeting agreed with our initial choices, each of the seven essays was picked as one of the best and as one of the worst by some of the teachers present. The results indicated that Diederich's "capricious judgment" was not restricted to high school and college English teachers. This was the beginning of an investigation into teacher judgment in the evaluation of written composition. Although further research is still in progress, the purpose of this article is to summarize the development of a test for measuring teacher judgment in evaluating themes and to summarize the results of the administration of that test to a sample of almost 200 teachers from grades one through twelve.

Composition Rating Scales

Although this investigation is not directly concerned with composition rating scales, a report by Follman and Anderson comparing five evaluation scales clarified the relation between teacher judgment and the use of such scales.[6] Their study compared four formal procedures and the "Everyman's Scale," an informal procedure by which the rater is instructed to use his own judgment in rating a set of themes. The di-

[6] Follman and Anderson, *op. cit.*

rections for the "Everyman's Scale" include the following paragraph:

> There is no particular grade that each essay should receive. You evaluate each essay according to *your own judgment* as to what constitutes writing ability. Use *your own judgment* about the writing ability as indicated by each essay. Don't use any system other than *your own judgment*.[7] (italics mine)

The results of the Follman-Anderson study indicate that the "Everyman's Scale" had unexpectedly high reliability coefficients (second highest of the five methods). The authors repeat the last three sentences from the paragraph quoted above and then add this comment.

> . . . A reasonable expectation is that such instructions would permit great individual difference and inconsistency among the raters using the *Everyman's Scale*. This did not occur; in fact, the opposite did.
>
> It may now be suggested that the unreliability usually obtained in the evaluation of essays occurs primarily because raters are to a considerable degree heterogeneous in academic background and have had different experiential backgrounds which are likely to produce different attitudes and values which operate significantly in their evaluation of essays. The function of a theme evaluation procedure, then, becomes that of a sensitizer or organizer of the rater's perception and gives direction to his attitudes and values; in other words, it points out what he should look for and guides *his judgment*.[8] (italics mine)

The Follman-Anderson research suggests the central role that teacher judgment plays in rating themes even when an evaluation scale is employed.

THE PROBLEM

This investigation is concerned with basic knowledge about teacher judgment as it exists and operates in the evaluation of themes in elementary and secondary classrooms. As an exploratory effort to measure and examine teacher judgment in this area, this study was designed to produce tentative answers to the following questions:

1. Can judgment in the evaluation of written composition be measured validly, efficiently, and reliably?
2. Is there agreement in judgment among experts as defined herein?

---

[7] *Ibid.*, p. 195.
[8] *Ibid.*, p. 198-199.

3. To what extent is there agreement in judgment among teachers of composition at three levels: elementary, junior high, and senior high? Since the high school and junior high teachers in the study have considerably greater academic background in English than the elementary teachers, another question is included in this one—Is academic background in English a factor in judgment?

4. How does the judgment of teachers at these three levels compare with that of experts?

5. How does the judgment of prospective and beginning secondary English teachers compare with that of the experts and with that of experienced secondary English teachers? Is teaching experience a factor in judgment?

6. How does the judgment of a select group of nonteachers compare with that of the experts and with that of secondary English teachers? Since the nonteachers in the study had academic backgrounds in English that were similar to those of the secondary English teachers (most of the nonteachers being English or journalism majors), this question also explores the possibility of a factor in judgment related to methodology or teacher education.

7. Are there teachers in any of the groups whose judgments are contrary to that of the experts and to that of the majority of other teachers?

Questions three through six were rewritten as twelve null hypotheses for testing.

METHOD

Test Development

The idea for the test originated with some work with the STEP Essay Test[9] in a pilot testing program in several elementary schools. Several samples of student written responses on the STEP Essay Test were used to screen and train lay readers, some of whom were to read and score the STEP essays. To make the selection of readers more objective, a scoring system was used based on deviations from the scores assigned by a small group of classroom teachers according to the suggested method for scoring these tests.[10]

Since the students wrote for 30 minutes on this test, these samples were relatively long, and scoring by reader applicants

---

9 *Sequential tests of educational progress, essay test, form 4A* (Princeton, N.J.: ETS, 1957).

10 *STEP handbook for essay tests, level 4* (Princeton, N.J.: ETS, 1957).

took some time. The 12 samples used originally were reduced to seven and later to five by eliminating those that had little discriminatory power, i.e., those that almost everyone agreed upon.

While the instrument was still in this preliminary stage, a variation of it was used in some inservice meetings with secondary English teachers. Some excerpts from the seven and later from the five samples were duplicated and given to teachers who were asked to pick the two best and two worst. Then the teachers were given the opinions of the elementary teachers on these samples, and a lively discussion usually followed.

Eventually some other shorter samples of writing from some other fifth grade classrooms were collected, and out of the new samples plus excerpts from the original samples two forms of the test were developed.[11] Each form of the test consists of five samples of writing which are to be ranked from best to worst. When either form is given to a group of 30 to 50 teachers, each sample is usually ranked as best by some and as worst by some.

The Scoring System

Each form of the 5-item test now consists of two essays that are ranked high (better) by a majority of teachers and two that are ranked low (worse) by a majority and one that is in between. After experimenting with several complex methods of scoring, none of which was satisfactory, the writer developed a simple easy-to-score system.

When either of the two "better" essays is ranked first or second, it scores one point. When either of the two "worse" essays is ranked fourth or fifth, it scores one point. The middle essay scores one point if it is ranked second, third, or fourth. Possible scores range from zero to five.

The Sample Population

The sample population included over 200 subjects who came from three sources: classroom teachers in the Jefferson County, Colorado, Public Schools; students in undergraduate and graduate classes in English and education at the University of Colorado; and a select group of nonteachers who were composition readers or composition reader applicants in this same school district.

---

[11] Form A of the test is included as an appendix to this article. Copyright, 1966, Vernon H. Smith.

The distribution of the sample population was as follows:

High School English Teachers	54
Junior High School English Teachers	44
Elementary Teachers (Grades 1-6)	32
Prospective and Beginning Secondary English Teachers	61
Nonteachers	27
Total	218

The experts in the study were five secondary English teachers who had been formally recognized as outstanding in the teaching of composition within their school districts or by some outside agency.

The groups that served as subjects in the development of the test, the sample population, and the experts were mutually exclusive.

Administration of the Test

The test was administered to the sample population in various groups. The experts took the test individually, usually by mail.

The directions are designed to be self-explanatory and to give no information that might bias the testee in ranking the essays. The following directions appear on each test:

> Below are five themes written by students in the same class. Using your usual criteria for evaluating written work, rank the five selections in order. Put a *1* in the blank to the right of the composition that you consider best, a *2* by the second best and so forth on to *5* which would indicate the worst.

The administration time for either form of the test is from 10 to 20 minutes.

In the testing situation nothing was said or discussed that would affect the results of a second administration of the test.

Statistical Procedures

The significance of the agreement of the five experts was determined by Snedecor's formula for intraclass correlation, an application of analysis of variance for a small group of raters recommended by Ebel in a study of various formulas for intraclass correlation.[12]

All other statistical procedures used can be found in standard statistics textbooks. The significance of differences among and between the subgroups in the sample population and the experts was determined by analysis of variance. The Pearson product-moment method was used to determine the test-

[12] R. L. Ebel, "Estimation of the reliability of ratings," *Psychometrika*, 1951, *16*, 407-424.

retest reliability coefficients and the reliability coefficient between the two forms of the test.

RESULTS

Validity

The relevance of the test is based on the assumption that the judgment used in ranking the five essays is the same or similar to the judgment used by teachers in evaluating students' written compositions. To the extent that this assumption is true, and only to that extent, the test has logical relevance. Whether this assumption is true or not, it is a common assumption underlying attempts to make the evaluation of essay tests and compositions more consistent and reliable.

The construct validity of the test is based on the agreement and consistency among the five experts. It was hypothesized that if the test were valid, experts in the teaching of composition would agree in ranking the essays, and that their judgment on two administrations of the test would be stable.

The experts took both forms of the test twice with an interval of six to ten weeks between administrations. The reliability of the scores of the experts is an essential part of the validity of the test.

Using Snedecor's formula for intraclass correlation,[13] the interrater reliabilities for the experts on Form A first and second administrations were .920 and .850, respectively, with reliabilities of average ratings .983 and .966. For Form B the interrater reliabilities were .880 and .840 with reliabilities of average ratings of .973 and .963. The coefficient of stability, test-retest reliability, for the experts on each form of the test was 1.00.

Reliability

The reliability of the test was determined by administering both forms of the test to two selected groups twice. The groups selected were students in two summer school classes at the University of Colorado, one in Teaching Reading in the Secondary School, the other in Teaching Literature to Adolescents. These two classes were selected because they enrolled a number of English teachers and because they were not courses that would intentionally affect the teachers' evaluation of written composition. The two test sessions were four weeks and two days apart. The test was not discussed in either class before or after either session. The students were told only that they were assisting in a study.

The coefficient of equivalence between Forms A and B

---

[13] *Ibid.*, p. 411.

was determined by using the results from the first administration of both forms. The coefficient of equivalence between forms was .607. The coefficients of stability for the two forms were determined by the test-retest method. The coefficients of stability were .739 and .790 for Forms A and B respectively. A third type of reliability coefficient, the coefficient of equivalence and stability, was determined by finding the correlation between the first administration of Form A and the second administration of Form B. The resulting coefficient was .679.

While reliability coefficients from .60 to .79 would not be highly regarded in the field of objective testing, they would certainly be considered significantly different than zero. This test attempted to measure judgment, a subjective factor, and these coefficients compare favorably with other studies of rater reliability.

There are two possible explanations for these relatively low reliabilities. Teacher judgment in the evaluation of written composition, the factor which the test attempted to measure, may not be very reliable. Diederich gives an example of reading a set of papers after an interval and finding a correlation of only .54 with his own first reading.[14]

The other possible explanation is that in creating a relatively simple, short instrument, length which might have improved reliability has been sacrificed. To test the latter a fourth reliability coefficient was calculated. The two forms were considered as one longer test of 10 items, and the test-retest reliability, the coefficient of stability, was .870. Therefore, the most reliable results will be obtained when the two forms of the test are given at the same time and the scores on each are combined to give a total score.

Results of the Sample Population

The consensus of all of the teachers in the sample and of all of the persons in each subgroup agreed with the consensus of the experts on the rank for each item on both forms of the test. However, there was much greater variance among the individual subjects in the sample population than among the experts. Analysis of variance indicated that all teachers in the sample and each of the five subgroups—nonteachers, elementary teachers, junior high English teachers, high school English teachers, and prospective and beginning teachers—differed significantly (at the .025 or .01 levels) from the experts. There

[14] P. B. Diederich, *op. cit.*, p. 6.

were no significant differences (.05 level) among the five subgroups.

Interpretation of Individual Test Scores

Individual test scores on Form A or B were interpreted as follows: (1) A score of *4* or *5* indicates agreement with the experts in judgment in the evaluation of compositions as measured by the test. Since the consensus of all teachers in the study agreed with the experts, the same scores indicate agreement with the consensus of composition teachers at all levels, grades one through twelve. (2) A score of *0, 1,* or *2,* indicates judgment that is contrary to that of the experts and to the consensus of composition teachers at all levels. To get a score of *2,* an individual has ranked three of the five themes in positions contrary to the ranking of the experts. (3) A score of *3* falls between the two extremes and indicates borderline or marginal agreement.

Interpretation of Scores of All Teachers

When the above interpretation of individual scores was used, almost half (48% on Form A; 47% on Form B) of the teachers in the sample agreed with the experts. However, more than half did not agree. On each form approximately 52% disagree or are borderline. Twelve per cent on Form A and 19% on Form B disagree or have judgment that is contrary to that of the experts. If the judgment of the experts as defined and measured in this study is accepted, then these persons are not competent to make such judgments. These results indicate an unpleasant situation for the child as he learns to write. His chances of getting a teacher whose judgment does not agree with the judgment of the experts are slightly greater than even. The chance that he will get a teacher whose judgment is contrary to that of the experts and to the consensus of other teachers of composition is about one in six; that is, he might expect two such teachers in his public school years.

Interpreting the Subgroup Scores

All subgroups differed significantly from the experts. In each group there were more persons who agreed with the experts than who were borderline or disagreed, but the combined total of those whose judgment was borderline and those whose judgment was contrary was greater than the number agreeing with the experts.

There were no significant differences among the five subgroups. The variations within each subgroup were much greater than the variations among the five subgroups. A difference between elementary and secondary teachers would have suggested that academic preparation was a possible

factor in judgment as measured by the test. A difference between prospective and beginning secondary English teachers and experienced secondary English teachers would have suggested that experience was a possible factor in judgment. A difference between nonteachers and teachers would have suggested that professional training was a possible factor in judgment. Since none of these differences was found, judgment as measured by the test may be independent of experience, academic preparation, and professional training.

CONCLUSIONS

1. The test results indicate that the subjective judgment of teachers in evaluating a specific set of short written compositions can be measured, and the results can be treated statistically.

2. Among experts in the teaching of composition, agreement in judgment, as measured by this test, does exist and is reliable.

3. Judgment, as measured by this test, is not related to experience, academic background, or professional training.

4. Although the consensus of teachers on any one item on the test agrees with the judgment of the experts, more than half of the teachers do not agree with the experts in judgment as measured by this test.

5. A significant number, between 10 and 20%, of classroom teachers charged with the responsibility for teaching students to write in grades one through twelve have judgment, as measured by this test, that is contrary to that of experts in the teaching of composition.

DISCUSSION

The development of this test and the subsequent research reported herein represent an exploratory probe into an area where little prior measurement had been attempted. The results reported here should be considered tentative until a body of research in this area becomes available.

There may be some questions about the nature and quality of the writing samples used in this study. The samples from the STEP Essay Test were used because they were available. Their use should not be regarded as an endorsement for the use of such mundane writing assignments by classroom teachers. Teacher judgment in the evaluation of more imaginative writing may be more complex and even more subjective than it was on the samples used in this test. In addition, the range of the five samples on any form of the test would not be at all typical of the range in any classroom. Obviously superior and

obviously poor samples could not be used in the test because they failed to discriminate well. The five samples on each form were picked on their discrimination value as test items, not on their literary merits.

The suggestion by Follman and Anderson that high reliability in the use of rating scales may be a measure of the homogeneity of the background of the raters rather than a measure of the scale poses a similar question with regard to this test.[15] Teachers in one large school district and undergraduate and graduate students in a few classes in one university could certainly be considered more homogeneous than a general population. Replication with a more heterogeneous population would be helpful.[16]

This writer expected to find differences in judgment due to experience, academic background, and professional training. The acceptance of the null hypotheses in these areas does not necessarily mean that such differences do not exist. They may exist, but they were not measured by this test.

This test should not be used to evaluate teacher competence in the teaching of composition. Too many other factors—including motivation, inspiration, and teaching techniques—play vital roles in teaching composition.

As students mature, essays and essay assignments are normally longer and more complex. Judgment on longer essays may be subject to greater variance than judgment on the relatively short samples in this test. This may account for the greater variance usually found among raters on themes at the college level.

Wright and Rubenstein found that poor writers had little ability to discriminate among compositions of varying merit.[17] In the same study, rank order assigned by good writers was close to that assigned by faculty members. These results suggest that judgment of written composition and writing ability may be related. This test could be used to determine the cor-

---

[15] Follman and Anderson, *op. cit.*

[16] The author wishes to encourage replication and further research. He will provide additional information, additional forms of the test as they are developed, and scoring instructions. Interested persons should write to Vernon H. Smith, School of Education, Indiana University, Bloomington, Indiana 47401.

[17] R. L. Wright and H. Rubenstein, "Can college students recognize good writing?" (Michigan State) *College of Education Quarterly*, 1960, *6*, 11-20.

relation, if any, between judgment and writing ability by giving the test to a group of subjects from whom writing samples were collected. If a significant correlation were found, follow-up studies might determine the effects of training in either area on the other. If some type of instruction in evaluation produces an improvement in writing ability, both teachers and students would benefit.

While this study investigated judgment, it made no attempt to investigate changes in judgment. Change in judgment under specified conditions offers opportunity for further research.

It is impossible to discuss the results of this study without confirming the lack of basic research in the teaching of written composition and in the evaluation of written composition which was pointed out by Braddock, Lloyd-Jones, and Schoer in *Research in Written Composition*.[18]

APPLICATIONS

The test developed in this study could be used for the following purposes:

1. to provide individual teachers and prospective teachers with knowledge of their judgment in the evaluation of written compositions;
2. to focus attention on the problems of evaluating written compositions by providing groups of teachers (secondary English departments, elementary school faculties, workshops, inservice training sessions) with a means of comparing individual judgments with other judgments within the group and with an outside criterion;
3. as part of a battery of tests to screen composition reader applicants;
4. as a tool to screen raters in research when judgment in the evaluation of written compositions is a factor.

In addition, the measurement technique developed in this study might be applied to any area when subjective judgment is a factor in evaluation.

FORM A*

I.

Dear Pen Pal,

I am in the fifth grade this year. I think I'm a very lucky

---

[18] R. Braddock, R. Lloyd-Jones, and L. Schoer, *op. cit.*

*Directions for the test are given in the article above.

Copyright, 1966, Vernon H. Smith.

boy. I have sevrail pets. I'm joining 4-H with my horse this year.

There are four people in my family. I'm also lucky I'm in this family. We do many things. Yesterday we went up in the mountains to get peat-moss.

I go to school at Lincoln. Where do you go to school? I have been lucky with teachers.

In the summer we go water skiing and camping. We mostly go water skiing. Its a lot of fun. What do you do in the summer?

Your friend
Eric

II.

Dear Pen Pal,

I live in Denver, I like where I live. I go to Lincoln that is the name of my school. My name is Beverly. I would like know your name? I have one brother and no sisters. My monther works ate the Honeywell Plant and my dad workes at Dave Cooks. I am in the fifth grade. My teacher's name is Miss Jones. My princeabule is Mrs. Brown. On saterdays we clean the house and, on sundays we rest and mother and I fix the diner.

Your friend,
Beverly

III.

Dear Pen Pal,

My name is leonard. You do not know me. I live in Colorado. My age is 10 years old.

My family lives here with me, but my brother does'nt. He lives in Texas. He works at a rocket fuel plant, which is called Rocketdyne. My mother just started to work on Monday. My father is a teacher. He teaches 7th grade geography, 8th grade American history, and 9th grade civics. My sister is trying to get a job.

My school is called Lincoln. It is a very nice school. My teacher's name is Miss Jones. And we, that is the whole school, have the nicest principal in the whole school district. Her name is Mrs. Brown.

When there is no school, I just ride my bicycle and play. I live in an apartment. There's a swimming pool but the manager closed it up so we can't swim in it until the first of June. When school is out for the day, almost everytime you see me I'm eating popcorn or drinking root beer. Unless I'm doing something else. There is a playground too at the aparl-

ments. There's a slide, a merry go round, 4 swings, and a jungle jim.

Very truly,
Leonard

IV.

Dear Pen Pal,

I am a girl. I live in the United States. My state name is Colorado I live in a suberb of Denver. I have blond hair and blue eyes.

I have lots of pets I have a dog, cat, and bird. My dog is a big dog, she is a Siberen Huskey, she has lots of fur on her. When we pluck her we get bags and bags of fur. In Winter when snow falls we hook her to a sled with her harnes on her back and she is read to pull. Down the street my dog and I go. My cat is gray, she has had several litters of kittens when we take them to be sold we half to make sure she doesn't see us or else she'll know that were taking them away. If she see's she will jump in the car and when we get to the place where we sell them she'll go where ever we go with her kittens and we'll loose her. We have this cat she is about seven years old and she is called a Purshan cat. She is a good cat.

Your friend
Diana

V.

Dear Pen Pal,

My name is Donnie. I am ten years I am 4 ft. 5" I have green eyes. The physical teacher work with me last year, I had lots of fun. He took it easy on me, because I had heart trouble.

I have a Mother and Father, two brothers one sister. We have a dog it's name is Skippy. It is a girl, black and brown, it's paws are white. My mother and father have black hair. One of my brothers and I have red hair. Our school had over three hundred people in it. This year I have nine classes alltogether. I have a very nice teacher, her name is Miss Jones.

On week ends I go over to my friends house, we ride bicks, we play football, and we play with are dogs. Mike and I have fun together, we ride our bicks to the school, not just to school but all over in our blocks. One day I got a box of raisens, Mike and I split the box of raisens.

Your friend,
Donnie